语义社会网络的超网络模型构建及关键节点自动化识别方法研究*

张 磊 马 静 李丹丹 沈 洋
(南京航空航天大学经济与管理学院 南京 210016)

摘要: 【目的】通过对语义社会网络的建模, 讨论如何识别对舆论传播演化起核心作用的关键节点。【方法】引入超网络理论对微博语义社会网络进行理论建模, 使用情感本体以及 LDA 话题模型对数据实现节点量化, 提出超边排序算法对用户节点进行计算和排序从而获取关键节点。【结果】利用真实微博网络数据编程实现超网络模型的构建和量化, 通过结果分析证明本文的关键节点识别方法在实际应用场景中的有效性和准确性。【局限】关键节点识别方法的实时应用效果和对识别关键节点后如何有效引导和干预机制未能全面涉及。【结论】本文的关键节点识别方法能够挖掘出微博网络的关键节点, 为政府对网络舆情监管和引导提供一种解决方案, 减少负面内容和消极舆论对互联网健康发展的影响。

关键词: 超网络 语义社会网络 关键节点识别 LDA 模型 情感本体

分类号: C931 G35

1 引 言

随着信息技术的发展, 社交网络如微博、微信等已成为人们日常生活中传播信息的主要手段。语义社会网络是一种由语义信息节点以及社会关系构成的新型复杂网络[1], 已成为互联网时代网络舆论传播的主要载体。互联网所具备的开放性、便捷性特点使得网络舆论表达更加自由、多元和难以控制, 负面内容和消极舆论严重阻碍了互联网的健康发展。由于语义社会网络这样的网络结构中多拥有一个或者多个处于核心地位的节点, 对网络结构和功能具备更大的影响力, 即关键节点[2]。根据信息传播的二八定律, 一般数量非常少的关键节点却可以影响到网络中大部分节点。例如微博中最具影响力的大 V 所发布的微博能够迅速地传遍整个网络[3]。因此在对网络舆论传播研究中, 特别是对突发舆情事件的研究[4-6]多侧重于通过关键节点控制和引导谣言和负面舆论的传播, 因此本研究具有重大现实意义。

关键节点的识别研究起源于社会网络分析。国内外关于网络舆论中关键节点的识别研究主要从网络拓扑结构和传播动力学方面切入。基于网络结构的节点重要性排序方法主要从网络的局部属性[7]、全局属性[8-9]、路径[10]、位置[11]以及节点移除和收缩[12]等方面进行衡量。Klemm 等[13]提出集群动力学中节点的重要性由网络结构和集群动力学机制共同决定的观点; Aral 等[14]对 Facebook 中 130 万用户传播行为研究发现用户影响力受到年龄、性别、婚姻等因素的影响。综上, 本文认为节点重要性不仅受到网络拓扑结构特性的影响, 同时也受到网络传播机制以及节点自身特性的影响。

传统网络模型[4,11]多由单一属性节点组成(多为用户节点), 对于该节点包含的语义、情感等其他属性等涉猎较少[15]。特别是语义社会网络包含多种不同要素

通讯作者: 马静, ORCID: 0000-0001-8472-2518, E-mail: majing5525@126.com。

*本文系国家自然科学基金项目"基于演化本体的网络舆情自适应跟踪方法研究"(项目编号:71373123)、江苏高校哲学社会科学研究重点项目"基于超网络的江苏教育微博舆情多元意见演化模型及应用研究"(项目编号:2015ZDIXM007)和校基本科研业务费重大项目培育基金"基于'模型–数据双驱动'的复杂社会网络行为大数据分析方法研究"(项目编号:P201630X)的研究成果之一。

的复杂关联关系，单一节点的网络模型无法准确描述真实社会网络。而美国科学家 Nagurney 等[16-17]率先定义的超网络为："高于而又超于现存网络的网络"，适用于刻画具备多层结构、多级特征、多属性的真实社会网络以及网络之间的相互作用和影响[18]。

目前基于超网络理论的关键节点识别研究应用领域较广，Lin 等[19]将其运用在电磁兼容性问题上，评估电子系统中的关键节点；Deng[20]应用于人群重要度建模，评估不同人群在领域中的重要作用；武澎等[21]利用特征向量中心性对社交超网络节点信息交互综合能力进行评判；马宁等[15]率先提出在论坛应用场景下的社交、环境、心理和观点 4 层超网络模型。本文针对微博应用场景下用户行为容易受到话题和情感属性影响的特性对模型子网的内容和构建方式进行改进；同时采用人工总结论坛语料库语义信息的方式构建网络和实验验证，还处于起步阶段，同时人工判断的方法耗时长，因此本文创新性引入情感本体和LDA主题模型对改进的话题和情感子网进行自动化识别和计算，并提出相应的排序算法以适应大数据场景。

2　微博语义社会网络超网络建模

超网络环境下语义社会网络关键节点的自动化识别研究思想如图 1 所示，首先在总结微博传播和用户特性基础上，从社交、内容、话题和观点 4 个维度构建超网络模型，刻画语义社会网络舆论的形成和演化过程；借鉴情感分析[22]和话题分析[23]方法，提出基于情感本体以及 LDA 话题模型的超网络节点自动化量化方法，同时提出 HyperEdgeRank 算法对超边进行排序，识别关键用户节点，最后通过实际数据分析可行性。本文以微博社交网络为应用场景进行阐述。

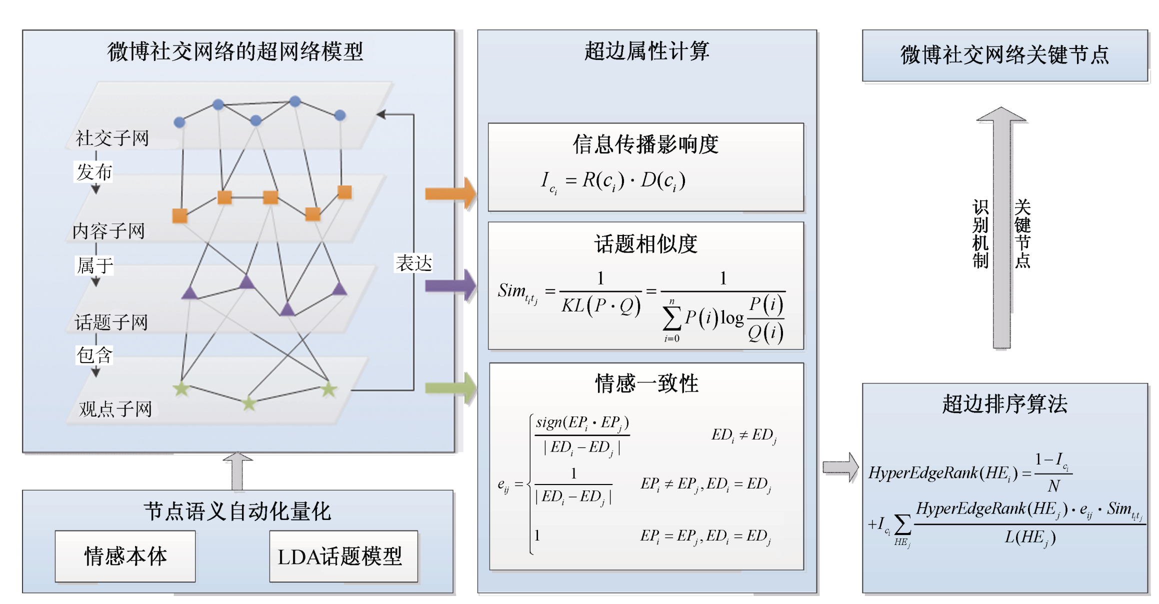


图 1　微博语义社会超网络关键节点自动化识别研究思想

2.1　超网络构建

在微博时代，每个人都是事件的传播者，微博以用户为中心，内容作为主体，网络工具为载体，向社会传播观点和信息[24]。因此在构建模型时，不仅要包括外在特征–社交主体用户，也应该囊括用户发布的内容以及其中所包含的话题和观点信息。本文的语义社会网络与传统研究的社交网络[15]的不同在于实时语义信息的引入，话题子网正是其语义信息的核心所在，而话题信息则是微博语义特征和内容信息的高度抽象；观点信息作为用户行为背后心理动机的抽象表达，对舆论的导向起到了主导作用，观点一般以情感表达。因此本文从网络属性和传播特征出发，改进前期研究的超网络模型[15]，从社交、内容、话题和观点 4 个层面构建超网络模型，提出了各子网内节点之间的关联关系，见图 1。

在微博社交网络中，社交主体(用户)利用内容(发布的微博)就某一话题表达观点，各子网的层内关系如下：

(1) 社交子网(Social Network): 以社交网络中参与讨论的社交主体即用户为节点, 用户之间的关注关系为边。

(2) 内容子网(Content Network): 以社交网络中用户发布的信息内容为节点, 微博之间存在的转发关系构建连边。

(3) 话题子网(Topic Network): 以从社交网络发布内容抽取的话题为节点, 包含相同关键词的微博话题的相似性关系构建连边。

(4) 观点子网(Emotion Network): 以从微博中提取的情感极性和情感强度作为节点, 具备相同的情感极性表明存在相关性, 构建连边。

超网络模型的层内关系是各子网络内要素之间的关系, 层间关系为各子网之间的关系。社交子网与内容子网的映射关系为用户节点对应多个微博内容节点, 来表征微博用户可以发布多条微博内容。内容子网与话题子网之间的映射关系为每条微博内容对应相关的话题。观点子网与话题子网的映射关系为用户在话题下的观点(本文设定为正面、负面和中立)。最后观点子网与社交子网之间的映射关系为用户发布微博参与某话题的观点倾向, 是隐性的映射关系。

超网络模型构建完成, 可用 $G=(V,HE)$ 表示, 其中 V 表示节点的集合, 即 $V=\{V_s,V_c,V_e,V_t \mid V_i \cap V_j \neq \varnothing\}$。HE 表示超边, 超边是 4 层不同子网节点之间的纵向连边, 表示用户 s_i 通过内容 c 就话题 t_k 发表观点 e_j, 用来表征不同类别节点之间的联系, 即层间关系, 如图 1 中 HE_1。

2.2 子网节点语义自动化量化方法设计

超网络模型前期研究[15,19-21]侧重于超网络模型的理论构建, 其中心理、观点这样的抽象属性难以利用定量的方式进行衡量, 多采用人工总结语料的识别方式。由于本文改进了微博语义社会网络的超网络模型, 对包含语义和情感抽象信息的话题子网和观点子网, 创新性地提出一种利用情感本体与 LDA 主题模型的自动化量化方法。

(1) 社交子网和内容子网都是从外在特征对社交网络的解读, 是微博传播模式的外在特征[7], 不具备明显的语义内涵, 因此直接通过数据集进行构建, 不量化。

(2) 为避免人工判别存在主观判断的影响, 引入 LDA(Latent Dirichlet Allocation)话题模型[23]对内容子网节点进行话题建模, 科学测量用户表达的语义内涵, 设置的话题数量即为话题子网的节点数量, 同时根据 LDA 的前提假设, 话题之间相互独立[23]。

LDA 概率话题模型是最常用的话题挖掘模型[23]。它的基本思想是假设每个文档为话题集的多项分布, 每个话题为所有词汇的多项分布, 将关键词–话题–文本的参数先验关系表达为三层贝叶斯模型。因此 LDA 话题抽取算法可根据关键词与话题的联合概率分布公式对已知微博文本和所有词汇进行重复抽样获取关键字与文本之间的共现概率, 推导获取文本与话题之间的联合概率分布, 从而实现话题节点的自动化抽取。本文通过 LDA 话题抽取算法[23]抽取 K 个话题, 将微博文本在话题集上的联合概率分布转化为微博内容与话题之间的对应关系, 实现内容子网与话题子网之间的关联。

(3) 借鉴情感分析的方法, 引入中文情感词汇本体库[22]抽取微博观点取向, 避免评判者的主观判断, 真实表征用户观点倾向。本文将微博经过中文分词之后, 利用情感词汇本体[22]进行极性标注, 累计情感词汇的情感强度和情感极性, 实现观点节点的抽取。

3 关键节点识别算法设计

3.1 超边排序算法(HyperEdgeRank)

本文采用超网络模型描述用户传播行为受到信息、话题以及情感等因素的影响, 因此与传统节点排序方法[11]对单一用户节点排序不同, 对超边进行排序, 将单一用户节点影响力计算转化为用户包含的所有超边影响力, 从而实现多维信息的综合考虑。同时本文研究目标是识别关键节点, 即该节点发布的内容对网络中其他节点产生巨大的影响。根据微博传播特性, 用户倾向于转发与自己观点一致的感兴趣话题的微博, 即容易受到这类用户节点的影响。因此本文认为某超边包含的微博节点的信息传播影响度越高, 即越多的用户可以接触到该微博信息, 那么该超边被其他超边链接的概率越大; 观点子网中某超边所含的观点类别与其他超边所含情感极性相同且情感强度相近, 话题子网中某超边的话题与其他超边的话题分布相似性越大, 该超边和其他超边链接获得的分值越大。因此在马宁等[15]研究的基础上, 从信息传播影响度、话

题相似度和观点一致性三个维度对超边排序算法迭代公式进行修改，得到：

$$HyperEdgeRank(HE_i)=\frac{1-I_{c_i}}{N}+I_{c_i}\sum_{HE_j}\frac{HyperEdgeRank(HE_j)\cdot e_{ij}\cdot Sim_{t_it_j}}{L(HE_j)} \quad (1)$$

其中，N 表示超边数，I_{c_i} 表示微博的信息传播影响度，e_{ij} 表示观点 e_i 和 e_j 之间的一致性，$Sim_{t_it_j}$ 表示话题 t_i 和 t_j 之间的相似性；$L(HE_j)$ 表示超边 HE_j 的超边连接度。

根据超网络的定义，本文引入两个超网络的属性指标：

(1) 节点超度(Node Hyperdegree)：表示包含该节点的超边数量[25]。

(2) 超边连接度(HyperEdge Degree)：超网络中，如果两条超边包含相同节点，说明两条超边通过该相同节点连接。超边连接度为超边通过所含节点与其他超边相连的超边数量[26]。

由于用户是微博内容的核心生产者和传播者，因此本文认为在利用超边排序算法对包含多维信息的超边进行排序后，仍然以用户节点为核心，累计社交子网中每一个用户节点参与的所有超边分值，通过与该节点的超度的比值获得用户节点的平均分值，分值最高的为关键节点，公式为：

$$Score(s_i)=\frac{\sum HyperEdge(HE_{s_i})}{HD_{s_i}} \quad (2)$$

3.2　超边子网属性计算

(1) 内容子网的微博信息传播影响度

内容子网中所有用户发布的一条微博代表一个微博信息节点 $c_i(1\leqslant i\leqslant n)$。微博内容在网络传播中影响到的用户数量越多，传播影响度则越高；微博内容被越多的人转发，传播影响度越高。由此可得信息传播影响度主要取决于传播的广度和深度。因此微博信息内容的传播影响度 I_{c_i} 主要取决于传播的广度和深度，引用马宁等[15]对信息传播影响度的定义，在微博语义社会网络中修正定义如下：

①信息传播广度 $R(c_i)$：微博信息节点的传播广度按照包含该节点的超边数 $P(c_i)$ 与总超边数 N 的比值进行衡量，即 $R(c_i)=\frac{P(c_i)}{N}$。

②信息传播深度 $D(c_i)$：微博信息传播的深度可理解为其经过转发后影响的用户数量，本文简化为微博信息节点 c_i 影响的社交子网中的用户数 $A(c_i)$，因此 $D(c_i)=\frac{P(c_i)/A(c_i)}{N/N_s}$，其中 N_s 表示社交子网中用户数。

由此得到信息传播影响度的公式为：

$$I_{c_i}=R(c_i)\cdot D(c_i)=\frac{P(c_i)^2\cdot N_s}{N^2\cdot A(c_i)} \quad (3)$$

(2) 话题子网的话题相似度

在计算话题节点之间相似度时引入统计自然语言常用的 Kullback-Leibler 距离度量[27]。由于 KL 距离越大表示话题之间的相似度越低，因此本文定义语义相似度 $Sim_{t_it_j}$ 表示话题节点 t_i 和 t_j 的相似度，与 KL 距离成反比，公式如下：

$$Sim_{t_it_j}=\frac{1}{KL(P\cdot Q)}=\frac{1}{\sum_{i=0}^{n}P(i)\log\frac{P(i)}{Q(i)}} \quad (4)$$

P 和 Q 分别为所有单词以 t_i 和 t_j 话题分布出现的事件。P(i) 表示第 i 个单词在话题 t_i 中出现的概率，Q(i) 表示第 i 个单词在话题 t_j 中出现的概率。由于 LDA 模型中每个话题向量是关于微博数据集包含的所有关键字的多项式分布，因此通过建模结果可获得 P(i) 和 Q(i)。

(3) 观点子网的观点一致性

对于同一话题不同的用户持有不同的情感极性，所发布的信息也具备不同的情感强度，因此观点节点包含不同的倾向和强度。本文创新性地利用情感本体获得信息节点的两个维度的情感信息：情感强度 ED_i 和情感极性 EP_i。$EP_i=1$ 时为正面观点，$EP_i=-1$ 时为负面观点，$EP_i=0$ 为中立观点。定义两个情感属性一致时，即情感极性一致且情感强度相近时，观点节点一致性更加明显。因此观点一致性 e_{ij} 由情感极性和情感强度共同决定且与情感强度的差值成反比，定义如下：

$$e_{ij}=\begin{cases}\frac{sign(EP_i\cdot EP_j)}{|ED_i-ED_j|} & ED_i\neq ED_j\\ \frac{1}{|ED_i-ED_j|} & EP_i\neq EP_j, ED_i=ED_j\\ 1 & EP_i=EP_j, ED_i=ED\end{cases} \quad (5)$$

其中，$sign(EP_i\cdot EP_j)$ 为符号函数，当 $EP_i\cdot EP_j>0$，

$\text{sign}(EP_i \cdot EP_j) = 1$ 表示情感极性相同；当 $EP_i \cdot EP_j \leqslant 0$，$\text{sign}(EP_i \cdot EP_j) = -1$ 表示情感极性相异。

4 验证与分析

4.1 数据处理

本文挖掘的节点为微博热门话题传播中对舆论导向产生重大影响的关键用户，因此在数据验证环节必须要基于热门话题数据。但是新浪微博自带的热门话题榜数据不开放 API 接口，因此只能在提取新浪热门话题榜前 5 个话题关键词："食品安全"、"贪腐"、"公务员考试"、"NBA"、"房价"的基础上，利用自行编写的爬虫程序通过微博移动客户端的搜索框抓取微博。在去除停用词后，由于过短的文本影响话题挖掘效果，因此筛除少于 20 个字的微博文本，总计获得 2014 年 4 月 30 日至 5 月 12 日的有效微博 526 条，参与用户共 429 人。采用开源的 NLPIR 分词和新词识别工具包[①]对所含微博文本在新词发现的基础上，实现分词，去除无语义内涵的高频词汇。同时利用大连理工大学信息检索实验室的中文情感词汇本体库[22]，借助 Java1.6 和 Matlab 编程工具实现超网络节点语义信息量化、各子网相似度计算和超边排序算法。

4.2 超网络模型的自动构建

(1) 话题子网节点量化结果

LDA 建模实验设置参数 $\alpha = 50/K$，$\beta = 0.01$，吉布斯采样的迭代次数为 1 000 次，由于 LDA 话题建模结果受到数据集和话题数量设置的影响，而本文的数据集较小，与数据来源(微博话题榜的前 5 个话题)保持一致，设置话题数量为 5。图 2 展示了 LDA 话题建模结果，由于 LDA 话题抽取结果是话题和词汇之间的联合分布，只能通过 Topic_i 进行表达。可以从话题的关键词集合中推断 Topic1 表征公务员考试话题，Topic2 表征 NBA 体育话题，Topic3 表征房价话题，Topic4 表征食品安全话题，Topic5 表征贪腐话题。由于爬虫程序是通过微博搜索框抓取的，LDA 模型结果显示抓取关键词与获取的话题信息相近，可见采用 LDA 模型的方式能够减轻人工判别话题信息的模糊性。

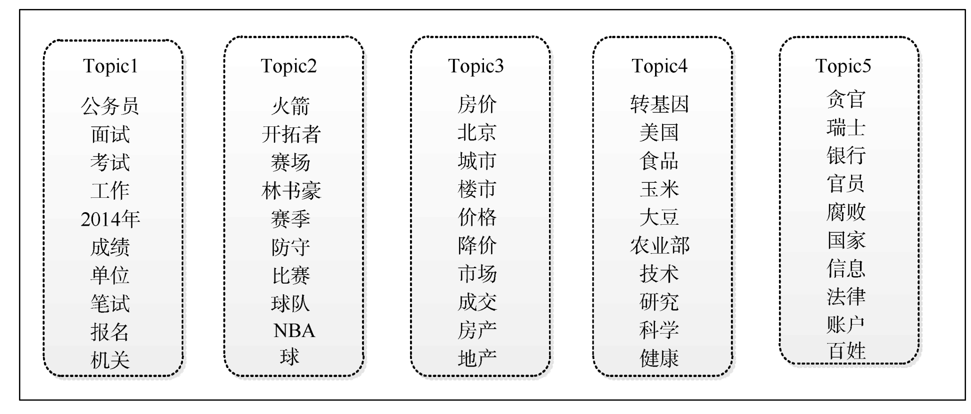


图 2 LDA 模型对话题子网节点量化结果

(2) 观点子网节点自动化量化结果

将内容子网中的信息节点，即微博文本，经过删除停用词预处理后获得候选信息文本。中文情感词汇本体[22]包含情感极性及情感强度，将情感分为 7 大类、20 小类，情感强度分为 1, 3, 5, 7, 9 等五档，9 表示强度最大。表 1 列出了内容子网微博节点与其抽取的观点节点的对应信息，本文的自动化识别方法通过情感强度和极性两个维度的测量保证了观点自动识别的准确性。

(3) 微博语义社会网络的超网络模型结果

在自动化量化观点和话题子网节点的基础上，构建超网络模型。本文的超网络模型中，社交子网中包含 429 个用户节点，内容子网中包含 526 个微博节点，观点子网中包含 64 个观点属性节点，话题子网中包含 5 个话题节点。表 2 展示了 4 层子网中各层节点的对应关系，即部分超边的组成情况。

①http: //ictclas.nlpir.org/docs.

表 1 内容子网节点与观点子网节点的对应表

微博节点	观点节点	情感极性	情感强度
c_1	e_{24}	1	2
c_2	e_{23}	–1	–19
c_3	e_{24}	1	2
c_4	e_{15}	–1	–15
c_5	e_1	0	0
…	…	…	…
c_{278}	e_{57}	1	6
c_{279}	e_{42}	–1	–3
c_{280}	e_{11}	–1	–13
c_{281}	e_{55}	1	5
c_{282}	e_{23}	–1	–19
…	…	…	…
c_{522}	e_{60}	–1	–7
c_{523}	e_{42}	–1	–3
c_{524}	e_{59}	1	7
c_{525}	e_{42}	–1	–3
c_{526}	e_{17}	–1	–16

4.3 超边排序算法结果

(1) 内容子网属性计算

超网络模型的内容子网共包含 526 个微博节点，按照公式(3)获得各信息节点的信息传播影响度，结果如表 3 所示(截取 15 个微博节点):

表 2 微博超网络模型部分超边组成

超边	社交子网	内容子网	话题子网	观点子网
HE_1	s_{365}	c_1	t_5	e_{24}
HE_2	s_{407}	c_2	t_1	e_{23}
HE_3	s_{48}	c_3	t_5	e_{24}
HE_4	s_{313}	c_4	t_4	e_{15}
HE_5	s_{73}	c_5	t_4	e_1
HE_6	s_{425}	c_6	t_5	e_{24}
HE_7	s_{272}	c_7	t_2	e_1
HE_8	s_{310}	c_8	t_4	e_{23}
HE_9	s_{110}	c_9	t_1	e_{58}
HE_{10}	s_{96}	c_{10}	t_2	e_1
…	…	…	…	…
HE_{517}	s_{13}	c_{517}	t_3	e_{25}
HE_{518}	s_{102}	c_{518}	t_4	e_6
HE_{519}	s_{68}	c_{519}	t_2	e_4
HE_{520}	s_{69}	c_{520}	t_1	e_1
HE_{521}	s_{73}	c_{521}	t_4	e_{60}
HE_{522}	s_{121}	c_{522}	t_3	e_{42}
HE_{523}	s_{241}	c_{523}	t_4	e_{59}
HE_{524}	s_{251}	c_{524}	t_3	e_{42}
HE_{525}	s_{287}	c_{525}	t_3	e_{17}
HE_{526}	s_{325}	c_{526}	t_4	e_{56}

表 3 信息传播影响度结果

c_i	$P(c_i)$	$A(c_i)$	N	$R(c_i)$	N_s	$D(c_i)$	I_{c_i}
1	1	13	526	0.001898	429	0.062619	0.000118821
2	2	11	526	0.003795	429	0.148008	0.000561699
3	1	14	526	0.001898	429	0.058146	0.000110334
4	1	5	526	0.001898	429	0.162808	0.000308934
5	7	3	526	0.013283	429	1.899431	0.02522963
225	1	11	526	0.001898	429	0.074004	0.000140425
226	1	9	526	0.001898	429	0.090449	0.00017163
227	5	13	526	0.009488	429	0.313093	0.002970522
228	1	12	526	0.001898	429	0.067837	0.000128723
229	1	4	526	0.001898	429	0.20351	0.000386168
255	1	7	526	0.001898	429	0.116292	0.000220667
256	1	10	526	0.001898	429	0.081404	0.000154467
257	1	12	526	0.001898	429	0.067837	0.000128723
258	1	14	526	0.001898	429	0.058146	0.000110334
259	1	1	526	0.001898	429	0.814042	0.001544671

(2) 观点子网属性计算

超网络模型共包含 64 个观点节点，根据公式(4)获得64个观点节点之间的相似度，结果如图3所示(截取前 10 个观点节点)。

526x526 double

	1	2	3	4	5	6	7	8	9	10
1	1	-0.0476	1	-0.0588	0	1	0	-0.0476	-0.1250	0
2	-0.0476	1	-0.0476	0.2500	0	-0.0476	0	1	0.0769	0
3	1	-0.0476	1	-0.0588	0	1	0	-0.0476	-0.1250	0
4	-0.0588	0.2500	-0.0588	1	0	-0.0588	0	0.2500	0.1111	0
5	0	0	0	0	1	0	1	0	0	1
6	1	-0.0476	1	-0.0588	0	1	0	-0.0476	-0.1250	0
7	0	0	0	0	1	0	1	0	0	1
8	-0.0476	1	-0.0476	0.2500	0	-0.0476	0	1	0.0769	0
9	-0.1250	0.0769	-0.1250	0.1111	0	-0.1250	0	0.0769	1	0
10	0	0	0	0	1	0	1	0	0	1

图 3　观点一致性计算结果截选

(3) 话题子网属性计算

超网络模型共包含 5 个话题节点，根据公式(4)获得 5 个话题节点之间的相似度，结果如表 4 所示：

表 4　话题相似度计算结果

	t_1	t_2	t_3	t_4	t_5
t_1	1	0.17242	0.18044	0.1865	0.16985
t_2	0.29845	1	0.26476	0.30823	0.29011
t_3	0.16008	0.14008	1	0.14578	0.13912
t_4	0.1686	0.15552	0.16177	1	0.15149
t_5	0.08366	0.10953	0.08451	0.08687	1

4.4　关键节点识别

借助 Matlab 实现超边排序算法，从而对所有超边进行计算，得到该模型 526 条超边的分值，截取排名前 19 条超边，结果如表 5 所示。

按照公式(2)对所有用户节点进行计算，获得 10 个关键用户节点和该节点超边平均值，分别为 s_{159}(0.19188)、s_{195}(0.19090)、s_{338}(0.19090)、s_{132}(0.19046)、s_{19}(0.19042)、s_{189}(0.19033)、s_6(0.19032)、s_{164}(0.19027)、s_{87}(0.19026)、s_{173}(0.19024)。

表 5　超边排序部分结果

超边排名	HyperEdge-Rank 值	超边编号	超边排名	HyperEdge-Rank 值	超边编号
1	0.191884	HE_{488}	11	0.190142	HE_{470}
2	0.190901	HE_{35}	12	0.190111	HE_{433}
3	0.190901	HE_{408}	13	0.190096	HE_{180}
4	0.190468	HE_{396}	14	0.190096	HE_{225}
5	0.190429	HE_{75}	15	0.190096	HE_{314}
6	0.190339	HE_{135}	16	0.190094	HE_{27}
7	0.190339	HE_{311}	17	0.190094	HE_{158}
8	0.190311	HE_{473}	18	0.190093	HE_{480}
9	0.190275	HE_{241}	19	0.190093	HE_{22}
10	0.190226	HE_{500}	…	…	…

根据表 6 截取的每个话题下分值较高的超边信息可以看出用户对于不同话题节点包含不同的情感倾向，对于公务员考试更多地持有中立的观点，对房价、食品安全以及贪腐话题相对持有负面观点占大多数，对于 NBA 体育话题正面观点是主流。实验结果与实际情况较为相符，可见基于超网络环境下的网络建模相比传统模型能够具备显示多维度、多层级、多属性信息的明显优势。

表 6　话题内超边分布情况(截取)

话题	用户节点	微博内容	EP_i	ED_i
Topic1	s_{69}	公务员考试考的是情商啊，逻辑思维，表达能力，应变能力，发散性性思维什么都考了。	0	0
Topic2	s_{68}	不是每个人都能成为詹韦杜科，但林书豪贝弗利这样的故事激励人们去努力，普通人的精神！	1	3
Topic3	s_{13}	唉，那些买了北京房子，持有北京房子的人们，到底在等什么呢？继续对政府抱有幻想？钱是一回事，生活品质是另一回事。到底我们需要的是钱还是品质？[可怜]就这样在雾霾中，被奴役下去吗？北京到底是谁的？[泪]//@张大伟 113：平均跌超过 5%//@古堡剑影：我询问了我家门口的房屋中介，二手房价确实略有下降	−1	−1
Topic4	s_{19}	此贴充分暴露了绅士明显在误导公众舆论，混淆相关性和因果性的区别//@崔永元：请中国科学家研究。//@annie 陈薇西：请相关部门重视此项研究成果并启动食品安全方面的调查！[围观]	−1	−5
Topic5	s_{55}	台湾的民主纵有万般谬误，能够生存至今已经打破了某些人所谓“民主不适合中国”的歪理邪说。民主制度下，有几个人胆敢像今日中国的贪官污吏那样前腐后继？中国的腐败成本占多少？三峡大坝土方工程中标价 60 元一方，最后一包 6 元。不出 20000 多条裂缝才怪呢。哪天三峡大坝垮了，五毛们就不喊了。	−1	−7

由于早期的网络科学研究关注的网络节点数目较少，可以通过问卷调查等方式以实际调查结果作为标准与其他算法结果进行比较和评价。但是大数据时代的来临，网络规模得到迅速增长，因此制定较为客观的节点重要性评价标准极为困难[2]。目前基于超网络理论的关键节点识别研究[6,15,19-21]评价各种算法优劣的主要思路是：以算法得出的重要节点作为研究对象，考察这些节点对整体网络结构和功能以及其他节点状态的影响程度来判断优劣。

表 7 中展示了本文数据集中前三位关键用户节点的超边组成情况，包括发表的信息内容、观点倾向以及话题内容。可见微博数据集中用户讨论的核心话题为 Topic4 食品安全问题，关键用户主要的观点倾向为负面。根据图 2 可知，话题关键词"转基因、美国、食品、中国"等都在表 7 中频繁出现。从前三位关键用户微博内容中也可以看出公众对于食品安全的负面情绪较强，特别是关键节点 s_{159} 观点的情感强度为 6，根据结果用户节点 s_{159} 比 s_{132} 在微博话题的传播过程中能够影响更多的节点，更为关键。单独考察这两个节点数据发现，节点 s_{159} 发表的微博内容影响 73 个用户人次对于食品安全话题的观点，而节点 s_{132} 发表的微博内容仅影响 24 个用户人次。虽然两者同为负面观点，但是前者具有更高的情感强度，对话题的引导性更强，也符合实验结果。可见本文的识别方法在实际应用中能够有效识别对舆论观点导向具有领导作用的关键节点。

表 7　关键节点用户超边组成情况

用户节点	微博内容	EP_i	ED_i	话题
s_{159}	中国大量进口转基因大豆就是从加入世贸后不久开始的，一直怀疑进口转基因与加入世贸有关。为加入世贸，中国接受了很多不公平的条件，比如在金融、贸易、投资方面的不对等开放。是谁把加入世贸的协定搞成了不平等条约，应该追究其责任。买办是逃脱不了历史的惩罚的，无论他多么会大义凛然的表演。	–1	–6	t_4
s_{195}	我对转基因食品最大的担忧就是不知道有什么样的潜在危险。做一件事情，不知道后果是什么，这是很可怕的。因为无法预知。	–1	–5	t_4
s_{338}	现在已经没有安全食品可食用了，地沟油，转基因，三聚氰胺，这让百姓还怎么活？	–1	–4	t_4

5　结　语

本文结合网络属性和微博传播机制，创新性地构建微博应用场景下的语义社会网络超网络模型，利用情感本体和 LDA 模型自动化构建观点和话题子网的语义节点，提出基于信息传播影响度、观点一致性和话题相似度计算方法，构建超边排序算法对超边进行计算和排序，计算社交子网中用户节点参与超边的平均累计分值实现关键节点的识别。使用了能够表征用户、内容、话题和情感属性的超边对节点重要性进行衡量，转变传统使用单一用户节点计算的局限性。最后通过实际数据验证了在语义社会网络关键节点识别中超网络理论的实用性以及超边排序算法的有效性，为舆情的监管和引导提供了一定的理论指导和解决方法。下一步工作是提高超网络节点自动化量化方法在大数据环境下的效率问题，增强对语义社会网络的实时监测效果；抛弃传统的删帖、禁言策略，研究在识别关键节点后如何进行有效的引导和干预。

参考文献：

[1] 辛宇，杨静，谢志强．基于随机游走的语义重叠社区发现算法[J]．计算机研究与发展，2015, 52(2): 499-511. (Xin Yu, Yang Jing, Xie Zhiqiang. A Semantic Overlapping Community Detecting Algorithm in Social Network Based on Random Walk [J]. Journal of Computer Research and Development, 2015, 52(2): 499-511.)

[2] 任晓龙，吕琳媛．网络重要节点排序方法综述[J]．科学通报，2014, 59(13): 1175-1197. (Ren Xiaolong, Lv Linyuan. Review of Ranking Nodes in Complex Networks [J]. Chinese Science Bulletin, 2014, 59(13): 1175-1197.)

[3] Weng J, Lim E P, Jiang J, et al. Twitter Rank: Finding Topic-sensitive Influential Twitterers[C]. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2010: 261-270.

[4] 康伟．基于 SNA 的突发事件网络舆情关键节点识别——以"7・23 动车事故"为例[J]．公共管理学报，2012, 9(3): 101-128. (Kang Wei. Analysis of the Key Nodes in Public Opinion Spread During Emergencies Based on Social Network Theory——A Case Study of the 7・23 Wenzhou High-speed

Train Collision [J]. Journal of Public Management, 2012, 9(3): 101-128.)

[5] 曹学艳，段飞飞，方宽，等．网络论坛视角下突发事件舆情的关键节点识别及分类研究[J]．图书情报工作，2014, 58(4): 65-70. (Cao Xueyan, Duan Feifei, Fang Kuan, et al. Research of Identification and Classification of Emergencies Key Nodes Based on BBS[J]. Library and Information Service, 2014, 58(4): 65-70.)

[6] 武澎，王恒山，李煜．突发事件信息传播超网络中枢纽节点的判定研究[J]．管理评论，2013, 25(6): 104-111. (Wu Peng, Wang Hengshan, Li Yu. Determination of the Hub Nodes in the Emergencies' Information Dissemination Supernetwork [J]. Management Review, 2013, 25(6): 104-111.)

[7] Bonacich P. Factoring and Weighting Approaches to Status Scores and Clique Identification [J]. The Journal of Mathematical Sociology, 1972, 2(1): 113-120.

[8] Zhang Z K, Zhou T, Zhang Y C. Tag-aware Recommender Systems: A State-of-the-art Survey[J]. Journal of Computer Science and Technology, 2011, 26(5): 767-777.

[9] 赫南，李德毅，淦文燕，等．复杂网络中重要性节点发掘综述[J]．计算机科学，2007, 34(12): 1-5, 17. (He Nan, Li Deyi, Gan Wenyan, et al. Mining Vital Nodes in Complex Networks [J]. Computer Science, 2007, 34(12): 1-5, 17.)

[10] Dolev S, Elovici Y, Puzis R. Routing Betweenness Centrality [J]. Journal of the ACM, 2010, 57(4): Article No.25.

[11] Kitsak M, Gallos L K, Havlin S, et al. Identifying Influential Spreaders in Complex networks [J]. Nature Physics, 2010, 6(11): 888-893.

[12] 许进．一种研究系统的新方法——核与核度法[J]．系统工程与电子技术，1994(6): 1-10. (Xu Jin. A New Method of Studying System——System Core and Coritivity [J]. Systems Engineering and Electronics, 1994(6): 1-10.)

[13] Klemm K, Serrano M A, Eguiluz V M, et al. A Measure of Individual Role in Collective Dynamics [J]. Scientific Reports, 2012, 2(2). Article No. 292.

[14] Aral S, Walker D. Identifying Influential and Susceptible Members of Social Networks [J]. Science, 2012, 337(6092): 337-341.

[15] 马宁，刘怡君．基于超网络中超边排序算法的网络舆论领袖识别[J]．系统工程，2012, 31(9): 1-10. (Ma Ning, Liu Yijun. Identification of Public Opinion Leader Based on the SuperEdgeRank Algorithm in Hypernetwork [J]. Systems Engineering, 2012, 31(9): 1-10.)

[16] Nagurney A, Dong J. Supernetworks: Decision-making for the Information Age [M]. Edward, Elgar Publishing, Incorporated, 2002.

[17] Nagurney A. Supernetworks: An Introduction to the Concept and Its Applications with a Specific Focus on Knowledge Supernetworks [J]. International Journal of Knowledge Culture and Change Management, 2005(4): 1-16.

[18] 王志平，王众托．超网络理论及其应用[M]．北京：科学出版社，2008. (Wang Zhiping, Wang Zhongtuo. Hypernetwork Theory and Application [M]. Beijing: Science Press, 2008.)

[19] Lin J, Dai F, Li B C, et al. Electromagnetic Compatibility Supernetwork Modeling and Node Importance Evaluation [C]. In: Proceedings of the 5th International Conference on Intelligent Human-Machine Systems and Cybernetics. IEEE Conference Publications, 2013: 306-310.

[20] Deng Z. Application of Crowd Importance Modeling in Regional Development Based on Super Network [J]. Value Engineering, 2015, 13: 211-212.

[21] 武澎，王恒山．基于特征向量中心性的社交信息超网络中重要节点的评判[J]．情报理论与实践，2014, 37(5): 107-113. (Wu Peng, Wang Hengshan. Key Nodes in Social Information Hypernetwork Evaluation Based on Eigenvector Centrality [J]. Information Studies: Theory & Application, 2014, 37(5): 107-113.)

[22] 陈建美．中文情感词汇本体的构建及其应用[D]．大连：大连理工大学，2008. (Chen Jianmei. The Construction and Application of Chinese Emotion Word Ontology [D]. Dalian: Dalian University of Technology, 2008.)

[23] Blei D, Ng A, Jordan M, et al. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[24] 袁立庠．微博的传播模式与传播效果[J]．安徽师范大学学报：人文社会科学版，2011, 39(6): 678-683. (Yuan Liyang. Communication Modes and Effects of Twitter [J]. Journal of Anhui Normal University: Humanities and Social Sciences, 2011, 39(6): 678-683.)

[25] Gourab G, Vinko Z, Guido C, et al. Random Hypergraphs and Their Applications [J]. Physical Review E, 2009, 79(6): 853-857.

[26] Wang J W, Rong L L, Deng Q H, et al. Evolving Hyper-network Model [J]. The European Physical Journal B, 2010, 77(4): 493-498.

[27] Kullback S, Leibler R A. On Information and Sufficiency [J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86.

作者贡献声明:

张磊: 设计研究方案, 设计并进行实验, 论文撰写;
马静: 提出研究思路及研究方案, 全文修改定稿;
李丹丹: 数据的抓取和处理;
沈洋: 提出部分修改建议。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 张磊, 马静, 李丹丹, 沈洋. JAVAprogram.rar. LDA 建模 Java 程序.
[2] 张磊, 马静, 李丹丹, 沈洋. data.txt. 分词后的数据集.
[3] 张磊, 马静, 李丹丹, 沈洋. LDAresult.twords. LDA 结果数据.
[4] 张磊, 马静, 李丹丹, 沈洋. emotionresult.xlsx. 情感属性识别结果.
[5] 张磊, 马静, 李丹丹, 沈洋. modelresult.xlsx. 微博超网络模型数据.
[6] 张磊, 马静, 李丹丹, 沈洋. matlabprogram.mat. 超边排序算法 Matlab 程序.
[7] 张磊, 马静, 李丹丹, 沈洋. finalresult.xlsx. 超边排序算法结果.

收稿日期: 2015-10-08
收修改稿日期: 2015-12-30

Hypernetwork Model for Semantic Social Network and Automatic Identification of Key Nodes

Zhang Lei　Ma Jing　Li Dandan　Shen Yang
(College of Economic and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: [**Objective**] This study aims to identify the key nodes of public opinion spread and evolution based on the semantic social network model. [**Methods**] We first built model for Weibo semantic social network with the help of hypernetwork theory, and then used emotion Ontology and LDA model to quantify nodes. Finally, we established the hyper edge sorting algorithm to identify the key nodes. [**Results**] The proposed model could effectively and acturately quantify those nodes from real Weibo data. [**Limitations**] We did not explore the results of the proposed method's real-time performance, and new ways of leading the public opinion after identifying those key nodes. [**Conclusions**] This study provides a solution for the government to identify the key nodes in the social network systems, and then reduce the impacts of negative contents to the healthy development of the Internet.

Keywords: Hypernetwork　Semantic social network　Key node identification　LDA model　Emotion Ontology